

Robust Estimation of Tokamak Energy Confinement Scaling through Geodesic Least Squares Regression

G. Verdoolaege^{1,2}

¹Department of Applied Physics, Ghent University, B-9000 Ghent, Belgium

²Laboratory for Plasma Physics, Royal Military Academy (LPP-ERM/KMS), B-1000 Brussels, Belgium

Corresponding Author: geert.verdoolaege@ugent.be

Abstract:

We address the scaling of the energy confinement time in tokamaks based on regression analysis applied to the multi-machine H-mode database. We briefly summarize various difficulties arising in estimating the scaling, either in terms of engineering quantities or dimensionless parameters. We invoke geodesic least squares (GLS) regression, an optimization technique that was recently developed to handle challenging regression problems in a robust way. Finally, possible routes are indicated to further increase the confidence in the global confinement dependencies and the predictions for future machines.

1 Introduction

The standard scaling law for the global energy confinement in H-mode tokamak plasmas provides a guideline for machine design and planning of operational scenarios. In addition, it is used as a benchmark to assess the quality of the confinement in present experiments. The currently recommended form of the scaling law is IBP98(y,2), which was obtained from a multi-machine database assembled using careful selection criteria [1]. The scaling law essentially provides the energy confinement time in terms of a number of key plasma quantities. The free parameters in this expression, i.e. the exponents in the power law, were estimated using statistical regression analysis based on ordinary least squares. Since then, however, the database has been expanded and different data selection criteria have been used, as well as different regression models and estimation methods. Unfortunately, depending on these various decisions, the estimated regression parameters can vary considerably. There is thus substantial uncertainty in the general dependence of the confinement on plasma and machine conditions across multiple machines. As a result, the confinement enhancement factor $H_{98(y,2)}$ (i.e. the ratio of an observed confinement time to the prediction given by the IBP98(y,2) scaling law) might give misleading

results in some instances and the predicted confinement times for next-step devices may be unreliable.

Given the continuing essential role played by the confinement scaling law, activities aiming at updating the scaling law using recent data are of clear importance. In this paper, we concentrate on the role of the regression method used to estimate the scaling parameters. The motivation is that, in light of the complexity and uncertainty in the confinement database, a robust regression technique is required that returns consistent estimates. Specifically, we present results of the confinement scaling, both in terms of engineering and dimensionless parameters, obtained with a robust technique called *geodesic least squares regression* (GLS). GLS was recently developed in the context of fusion scaling laws and applied to estimation of the scaling for the L-H power threshold [2].

The outline of the paper is as follows. The main difficulties concerning estimation and interpretation of the confinement scaling law are discussed in Section 2. We then very briefly mention the basics of GLS regression in Section 3, followed by a presentation of our results of the application of GLS to the confinement data in Section 4. Section 5 concludes the paper and lists some opportunities for future work.

2 The IPB98(y,2) scaling law

The mathematical form of the IPB98(y,2) scaling law for the thermal energy confinement time $\tau_{E,th}$ (s) in ELMy H-mode tokamak plasmas is given as follows, in terms of engineering variables:

$$\tau_{E,th} = \beta_0 I_p^{\beta_I} B_t^{\beta_B} \bar{n}_e^{\beta_n} P_1^{\beta_P} R^{\beta_R} \kappa^{\beta_\kappa} \epsilon^{\beta_\epsilon} M_{eff}^{\beta_M}. \quad (1)$$

Here, I_p is the plasma current (MA), B_t the vacuum toroidal magnetic field (T), \bar{n}_e the central line-averaged electron density (10^{19} m^{-3}), P_1 the net input power (MW) (corrected for losses due to charge exchange with the heating beam and unconfined orbits), R the plasma major radius (m), $\epsilon = a/R$ the inverse aspect ratio (with a the plasma minor radius in m), κ the plasma elongation and M_{eff} the effective atomic mass. IPB98(y,2) was derived by fitting (1) to the so-called ‘standard set’ in the 1997 version DB3 of the global H-mode confinement database [1, 3, 4].¹ Usually the estimation is performed on a logarithmic scale, turning the problem formally into a linear regression analysis with known properties of the statistical estimators. Furthermore, in the case of IPB98(y,2) the Kadomtsev constraint was imposed, as well as a fixed exponent β_B for the magnetic field. The latter constraint was motivated by the observation that estimates of the exponents for B_t and I_p turn out to be linked, reflecting the poor condition of the data set w.r.t. the safety factor [5]. Consequently, a value of $\beta_B = 0.15$ was fixed, at the time roughly in line with scans in individual devices (see [6], pp. 317–318). Poor conditioning of the database w.r.t. κ was also noticed [5]. The estimates of the regression parameters in IPB98(y,2) are shown in Table I.

In later work, emphasis shifted to alternative statistical algorithms to perform the fit. Indeed, it was noticed that the error bar on the predictor variable P_1 (ca. 14%) is compa-

¹To be precise, version DB2.8 was used, including NBI discharges only (see [1], pp. 2207–2208).

TABLE I: REGRESSION ESTIMATES FOR THE IPB98(Y,2) SCALING LAW AND RESULTS OF OUR ANALYSIS USING OLS, MAP AND GLS (LINEAR ON THE LOGARITHMIC SCALE (‘LIN.’) AND NONLINEAR ON THE POWER LAW (‘POW.’)) BASED ON THE ENGINEERING FORM OF THE SCALING LAW.

		β_0	β_I	β_B	β_n	β_P	β_R	β_κ	β_ϵ	β_M	$\hat{\tau}_{E,th}$ (s)
IPB98		0.056	0.93	0.15	0.41	-0.69	1.97	0.78	0.58	0.19	4.9
OLS	Lin.	0.049	0.78	0.32	0.43	-0.67	2.22	0.38	0.57	0.18	4.2
	Pow.	0.055	0.76	0.42	0.41	-0.81	2.40	0.97	0.62	-0.22	3.3
MAP	Lin.	0.056	0.89	0.20	0.35	-0.59	1.94	0.20	0.36	0.17	4.1
	Pow.	0.053	0.88	0.21	0.37	-0.59	1.96	0.21	0.37	0.13	4.1
GLS	Lin.	0.054	0.73	0.37	0.44	-0.72	2.33	0.53	0.72	0.20	3.9
	Pow.	0.048	0.66	0.44	0.49	-0.74	2.48	0.64	0.84	0.19	4.0

rable to that on the response variable $\tau_{E,th}$ (ca. 20%) [3]. This is a violation of one of the main assumptions underlying OLS regression, where the predictor variables are taken to be infinitely precise quantities (or at least with error bars that are negligible compared to the uncertainty on the response variable). Therefore, an error-in-variables (EIV) technique was employed, based on a principal component analysis, and some differences in the scaling law estimates were observed, as compared to the OLS results [3]. However, it was also noticed that the exponents for \bar{n}_e , P_1 and R were rather sensitive to the assumed error bars on the plasma thermal energy W_{th} (where $\tau_{E,th} = W_{th}/P_1$) [3, 5]. Consequently, the EIV estimates were deemed unreliable and therefore the ’98 result remained the recommended scaling law. We argue here that a certain degree of arbitrariness in the EIV results does not make OLS a better alternative. Indeed, it is important to realize that, although OLS, due to its simplicity, yields a single estimate without any further degrees of freedom, many assumptions are implicit that can equally well invalidate its results [2].

In order to make the link of the confinement scaling with theory more clear, it is furthermore customary to express the scaling in terms of dimensionless parameters. We follow [5] in the choice of variables, yielding

$$\omega_{ci}\tau_{E,th} = \alpha_0 \rho^{*\alpha_\rho} \beta_t^{\alpha_\beta} \nu^{*\alpha_\nu} q_{95}^{\alpha_q} \kappa^{\alpha_\kappa} \epsilon^{\alpha_\epsilon} M_{eff}^{\alpha_M}. \quad (2)$$

In this expression, ω_{ci} is the ion gyro-frequency in units of 10^8 rad s^{-1} , ρ^* is the normalized ion Larmor radius, β_t the plasma pressure normalized by the toroidal field, ν^* the normalized collisionality and q_{95} the safety factor at 95% of the toroidal flux. These

variables can be calculated from the engineering variables as follows:

$$\begin{aligned}\rho^* &= \left(\frac{2m_p}{3e^2}\right)^{1/2} \left(\frac{M_{\text{eff}}W_{\text{th}}}{V\bar{n}_e}\right)^{1/2} \frac{1}{B_t\epsilon R}, & \beta &= \frac{4\mu_0}{3} \frac{W_{\text{th}}}{VB_t^2}, \\ \nu^* &= \frac{15e^4 \ln \Lambda_{\text{ei}}}{4\pi^{3/2}\epsilon_0^2\mu_0} \frac{V^2 R^2 B_t \epsilon^{1/2} \bar{n}_e^3 \kappa}{W_{\text{th}}^2 I_p}.\end{aligned}$$

Here, m_p is the proton mass, e the elementary charge, ϵ_0 and μ_0 the permittivity, resp. permeability of the vacuum, $\ln \Lambda_{\text{ei}}$ the Coulomb logarithm for electron-ion collisions, V the plasma volume and all variables are in SI units.

However, reliable estimation of the exponents in the dimensionless form of the scaling law is substantially more difficult than in the case of the engineering parameters. The variables ρ^* , β_t and ν^* are affected by considerable error bars, increasing the necessity for a statistical method that can properly handle uncertainty in all variables. Furthermore, concern has been raised about the collinearity of several dimensionless variables, in particular due to the correlation between β_t and ϵ in the database (low aspect ratio devices tended to operate at higher β_t) [7]. This further complicates the estimation. Some proposals have been made to substitute the dimensionless quantities by other variables that are less correlated and less affected by uncertainty [3, 6], but it is not clear to what extent this has really improved the situation (it is certainly fruitless under the OLS assumptions [8]).

In fact, it is common practice to derive the dimensionless form of the scaling law from the estimates obtained by performing regression analysis on the engineering expression, simply by transforming the exponents. The result for IPB98(y,2) is given in Table II. The scaling with ρ^* , β_t and ν^* is of particular interest. Indeed, most estimates for the ρ^* dependence are consistent with $\omega_{ci}\tau_{\text{E,th}} \sim \rho^{*-3}$, which has a known interpretation in terms of a gyro-Bohm scaling [1]. The negative scaling with β_t is unfavorable at high- β operation and has been the subject of much investigation [5]. Finally, the almost negligible dependence of confinement on ν^* suggested by the IPB98 scaling is not in line with a stronger negative scaling seen in dedicated scans on individual machines [5].

Direct estimation on the dimensionless form of the scaling law was done in [7] using a classic EIV analysis and a Bayesian procedure, but the estimates were deemed unreliable for prediction. In the present paper, we make another attempt using GLS regression.

3 Geodesic least squares regression

The idea behind OLS regression for a single response variable y is to estimate the parameters β_k ($k = 0, \dots, p$) of the regression model by minimizing the difference between, on the one hand, the prediction of the values of y , given n measurements x_{ij} of the m predictor variables x_j , and, on the other hand, the actually observed values y_i ($i = 1, \dots, n$, $j = 1, \dots, m$). However, this only takes into account the statistical error on y , whereas the x_j may have non-negligible uncertainty as well. This issue is also addressed by classic EIV methods based on principal component analysis, and it is especially important in the present context, when regressing directly on the dimensionless form (2) of the scaling law.

TABLE II: REGRESSION ESTIMATES FOR THE IPB98(Y,2) SCALING LAW IN DIMENSIONLESS FORM, TRANSLATED FROM THE EXPONENTS OF THE ENGINEERING PARAMETERS IN TABLE I, AS WELL AS THE RESULTS OF OUR ANALYSIS USING OLS, MAP AND GLS (LINEAR ON THE LOGARITHMIC SCALE (‘LIN.’) AND NONLINEAR ON THE POWER LAW (‘POW.’)) DIRECTLY ON THE DIMENSIONLESS QUANTITIES.

		α_0	α_ρ	α_β	α_ν	α_q	α_κ	α_ϵ	α_M
IPB98		7.21×10^{-8}	-2.70	-0.90	-0.01	-3.0	3.3	0.73	0.96
OLS	Lin.	1.36×10^{-7}	-2.45	0.36	-0.39	-0.48	0.89	-1.61	1.46
	Pow.	1.51×10^{-6}	-2.20	0.33	-0.43	-0.24	0.92	-0.45	1.12
MAP	Lin.	4.84×10^{-7}	-2.24	0.28	-0.39	-0.55	0.98	-1.34	1.15
	Pow.	1.33×10^{-6}	-2.06	0.23	-0.40	-0.55	1.01	-1.08	1.17
GLS	Lin.	1.86×10^{-8}	-2.91	0.64	-0.38	-0.35	0.50	-2.25	1.87
	Pow.	0.55×10^{-6}	-2.62	0.45	-0.43	-0.53	0.71	-1.55	1.46

Furthermore, the errors on the various quantities may be different from one measurement to another. In the present context, the relative errors on the quantities in the global H-mode database indeed differ between machines. A way around this is to consider the more general maximum likelihood method (ML), which maximizes the probability distribution of the response variable conditional on the predictor variables. For the remainder of the paper we will assume normally distributed uncertainties, reducing ML to the following optimization problem:

$$\{\hat{\beta}_k\} = \arg \max_{\{\beta_k\}} p_m,$$

$$p_m \equiv \frac{1}{\sqrt{2\pi}\sigma_m} \exp \left\{ - \sum_{i=1}^n \frac{[y_i - f(\{x_{ij}\}, \{\beta_k\})]^2}{2\sigma_m^2} \right\}.$$

Here, f is the regression function (possibly nonlinear), while the measurements y_i are assumed to be mutually independent, and similar for the x_{ij} . The standard deviation σ_m in general describes uncertainty on the response *and* the predictor variables. We refer to σ_m as the standard deviation of the *modeled distribution* p_m , since it depends on the regression model: the uncertainty on the x_j propagates through f . Incidentally, when suitable prior distributions are introduced on the β_k in a Bayesian framework, ML becomes the maximum *a posteriori* method (MAP) [2], which we will compare with in our regression experiments.

There is one flaw in this reasoning, which is shared by most regression methods, in-

cluding many of the more sophisticated. It assumes that σ_m is indeed the correct standard deviation on the y_i , leaving no room for unforeseen sources of uncertainty. Still, such additional uncertainties often occur, e.g. due to outliers in the data, plasma fluctuations or transients, uncertainty in the regression model, etc. The GLS regression method accommodates these situations by considering, apart from p_m , another distribution for the dependent variable that makes as few assumptions about the data as possible. We call this the *observed distribution* p_o , and here we will assume only that it is a normal distribution centered on each measurement y_i , with some unknown standard deviation σ_o that is to be estimated from the data. As such, every measurement y_i is actually treated as a probability distribution and GLS aims to minimize the overall difference between the modeled and observed distributions, just like OLS minimizes the overall difference between the modeled and observed values of y . As a distance measure between probability distributions we choose the *geodesic distance* (GD) rooted in information geometry, which is a geometric approach to probability theory [2]. Hence, the $p + 2$ parameters $\beta_0, \dots, \beta_p, \sigma_o$ are estimated through the following optimization problem:

$$\{\hat{\beta}_k, \hat{\sigma}_o\} = \arg \min_{\{\beta_k, \sigma_o\}} \text{GD}^2 \left[\prod_{i=1}^n p_o(y|y_i, \sigma_o), \prod_{i=1}^n p_m(y|\{x_{ij}\}, \{\beta_k\}) \right].$$

It has been demonstrated that, despite its simplicity, GLS consistently outperforms several other regression methods in various challenging regression tasks [2]. Nevertheless, several aspects of the GLS method are still open for improvement, as part of future work.

4 Confinement scaling

We now apply the GLS technique to estimate the confinement scaling law using the DB3 database. With the additional requirement to calculate the dimensionless parameters, the number of entries in the database was limited to 1296 from 9 devices. The database contains error estimates for each of the variables, in terms of percentages. We use these relative errors to derive the standard deviations that are required in the GLS method. As mentioned before, the error bar on a specific quantity, in particular the measured confinement time, can be different from one machine to another. Therefore, for each machine we need a parameter representing the observed standard deviation, which was considered as a relative error w.r.t. the measurements (observed mean) for each variable. We did not impose any constraints and the data were not weighted.

Some first results of the scaling with engineering parameters were presented in [9], where the influence of the constraint $\beta_B = 0.15$ was also studied. In Table I we repeat the analysis on the present database and, in addition to the comparison with OLS, we also mention the results of a regression analysis by means of MAP. As argued in [2, 9], it is worthwhile to carry out the analysis on the original power law, i.e. without taking the logarithm, as this may artificially influence the results. Hence, Table I mentions both the results using linear regression on the logarithmic scale and power-law regression. If all assumptions underlying OLS were fulfilled, the results would ideally be the same. That

this is not the case, in particular for the exponents related to the geometrical quantities and the atomic mass, as well as the predicted confinement time, indicates that OLS regression on the global H-mode database may be problematic. In contrast, the results obtained using MAP and GLS are much more consistent, whether log-linear or power-law regression is used, although there are still differences between the estimates of these two more advanced methods. Furthermore, the trade-off between β_I and β_B can be noticed, while the largest discrepancies between MAP and GLS occur again in the exponents for the geometrical parameters and M_{eff} . In addition, the confinement time predictions for ITER are given in the table, under the conditions $I_p = 15$ MA, $B_t = 5.3$ T, $\bar{n}_e = 10.3 \times 10^{19} \text{ m}^{-3}$, $P_l = 87$ MW, $R = 6.2$ m, $\kappa = 1.7$, $\epsilon = 0.32$ and $M = 2.5$. Error bars on all estimates will be given in a later, more comprehensive analysis.

Next, Table II contains the parameter estimates using OLS, MAP and GLS, obtained by direct regression on the dimensionless form (2) of the scaling law. It is noteworthy that the α_ρ estimates suggest a Bohm-like scaling ($\omega_{ci}\tau_{E,\text{th}} \sim \rho^{*-2}$), while the GLS results point somewhat more at gyro-Bohm scaling. Interestingly, the unfavorable scaling with β_t has been replaced by a slightly positive dependence in all regression estimates. However, in the light of the tendency of collinearity between β_t and ϵ , combined with a sign change of the ϵ scaling, this should be interpreted with great care. It is clear that further enhancements to the database could be very valuable in order to resolve this uncertainty. On the other hand, we aim to investigate in future work how GLS can be enhanced in order to better deal with situations of near-collinearity. Further results to be noted, compared to IPB98(y,2), are the stronger negative dependence on ν^* , as well as the much weaker dependence on κ , both of which are also captured by OLS and MAP.

5 Conclusion

We have addressed the energy confinement scaling law in tokamaks, based on regression analysis using data from the global H-mode database. The advantages of geodesic least squares regression were pointed out, leading to enhanced robustness of the method. This was demonstrated by regression analysis applied to the confinement scaling, on a logarithmic scale or with the original power-law model. With this work, we explicitly aim to increase the robustness and consistency between regression analysis on the scaling law in terms of engineering quantities or dimensionless variables. However, the initial results reported in this paper indicate that further work, both to improve the database conditioning as well as the robustness of the regression analysis, may be necessary to accomplish this goal. As far as GLS regression is concerned, we intend to improve the optimization algorithm used to find the minimum aggregated geodesic distance between the modeled and observed distributions. In addition, a Bayesian analysis returning a joint posterior distribution for the regression parameters would provide certain advantages compared to a pure optimization-based approach. Finally, visualization of the database using projection methods may contribute to identifying voids or excessive outliers in the database.

References

- [1] ITER Physics Basis 1999 *Nucl. Fusion* **39** 2175–2249
- [2] Verdoolaege G and Noterdaeme J M 2015 *Nucl. Fusion* **55**
- [3] ITER Confinement Database and Modelling Working Group 1997 *Plasma Phys. Control. Fusion* **39** B115–B127
- [4] Global H-mode confinement database DB3, online at <http://efdasql.ipp.mpg.de/hmodepublic>
- [5] McDonald D *et al.* 2007 *Nucl. Fusion* **47** 147–174
- [6] Christiansen J *et al.* 1992 *Nucl. Fusion* **32** 291–338
- [7] Kaye S *et al.* 2006 *Plasma Phys. Control. Fusion* **48** A429–A438
- [8] Chudnovskiy A 2006 *Plasma Phys. Rep.* **32** 344–348
- [9] Verdoolaege G, Shabbir A and Hornung G 2016 *Rev. Sci. Instrum.* **87**